Semantic Segmentation in the Task of Long-Term Visual Localization

Lukáš Bureš(✉) and Luděk Müller

Faculty of Applied Sciences, New Technologies for the Information Society, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic
{lbures,muller}@ntis.zcu.cz

Abstract. In this paper, it is discussed the problem of long-term visual localization with a using of the Aachen Day-Night dataset. Our experiments confirmed that carefully fine-tuning parameters of the Hierarchical Localization method can lead to enhance the visual localization accuracy. Next, our experiments show that it is possible to find an image's area that does not add any valuable information in long-term visual localization and can be removed without losing the localization accuracy. The approach of using the method of semantic segmentation for preprocessing helped to achieve comparable state-of-the-art results in the Aachen Day-Night dataset.

Keywords: Long-term visual localization · Semantic segmentation · SuperPoint · SuperGlue · HRNet-OCR

1 Introduction

The ability to effectively represent local visual information is the key to a vast range of computer vision applications. The range of these applications may contain image alignment, which requires local image descriptors to be accurately matched between different views of the same scene, image classification, and image retrieval where massive descriptor collections are frequently scanned to locate the ones most relevant to those of a query image. The recent summary can be found in [20].

At the core of the problem, there is a challenge of extracting local representations at locations of keypoints. The keypoints are typically distributed sparsely in the image in a discriminative and invariant manner to various image transformations. Additional requirements, often critical, are connected with an efficiency of a representation in terms of the computational costs required to produce it, the space required to store it, and the time required to search for matching descriptors in large descriptor repositories.

The problem of the long-term visual localization can be described as follows: estimating the 6 DoF camera pose from which a given image was taken relative to a reference scene representation. Visual localization is a key technology for augmented, mixed, and virtual reality, robotics, and self-driving cars.

© Springer Nature Switzerland AG 2021
A. Ronzhin et al. (Eds.): ICR 2021, LNAI 12998, pp. 27–39, 2021.
https://doi.org/10.1007/978-3-030-87725-5_3

To evaluate visual localization over time, a benchmark dataset is needed. It aims for evaluating 6 DoF pose estimation accuracy over large appearance variations caused by changes in seasonal and during different illumination conditions.

Long-term visual localization has multiple subtasks in the pipeline where the improvements to state-of-the-art can be made. A few open areas where is always space for improvement can be mentioned.

The first step in the pipeline it is usually the detecting of keypoints. End-to-end trained neural networks replaced the traditional approaches in which corners and similar points were considered to be keypoints. During the training process, there are multiple ways where the neural network can be improved, e.g., training data, the structure of the neural network, training procedure itself, optimizing function.

The next part is a keypoint descriptors which is used later for matching purposes. In traditional methods, the keypoint descriptor was represented by a feature vector describing a local window around the detected keypoint. The current describing approaches can be replaced by a neural network trained in an end-to-end manner. There are similar options, like in the case of keypoint detector, how to improve descriptor. However, the ultimate goal here is to get the most general descriptor unbiased to the given image.

The third area of improvement can be made in a keypoint matcher. Originally keypoint matchers like brute force or nearest neighbor were used. In the current state-of-the-art methods, the used matchers are based on neural networks which mostly outperform traditional matchers with a significant margin. The matcher has to be robust and match the right keypoints even if there are multiple instances of the same objects, e.g., tiles, windows, and fence.

Before the matching phase even starts, coarse matching is mostly used for speeding the process. The coarse matching helps to find the N most visually similar images and order them by a similarity value. It helps tremendously because the matching stage is not needed to match every possible query-dataset image combination exhaustively. Improving the global descriptor to find the most similar image to a given query image can help improve long-term visual localization overall.

Structure from motion (SfM) is also a part of the pipeline. It can be improved in a way, e.g., improve the number of keypoint co-occurrences, improve the robustness of co-occurrences or use a different type of data.

A carefully fine-tuning of a parameter of a method can improve the results of a given method. The importance is to do so without overlearning the training set and the test set.

Before the pipeline even starts to detect keypoints and process dataset, a preprocessing can be applied for enhancing input images. A similar approach can be applied to postprocessing images or keypoint descriptors before the data goes to the next stage of the pipeline.

This paper focuses on the last two points fine-tuning parameters of a method and preprocessing data.

2 Dataset

For the experiments, the Aachen Day-Night dataset was used [12]. It is a subset of the original Aachen dataset [13] which contains 1.5M 3D points and 369 query images. This Aachen Day-Night dataset consists of 4328 reference images and 922 query images. In sum, there are 824 images captured in the daytime and 98 ones taken in the nighttime.

The entire dataset was captured in the old inner city of Aachen, Germany, by hand-held cameras and mobile phones over about two years. All of the query images were captured during the nighttime and daytime and only using mobile phone cameras. Therefore, this dataset is suitable for solving localization problems using mobile devices, e.g., Augmented or Mixed Reality. The 98 nighttime query images were taken by using software HDR to obtain high-quality and well-illuminated images.

The collection of images was publicly released, and it is not intended for commercial use and should be applied primarily for research purposes.

The Aachen Day-Night dataset (see example in Fig. 1) was used to prove that the proposed Structure from motion using dense convolutional neural network features with the keypoint relocalization outperforms a state-of-the-art Structure from motion (COLMAP[1] [14,16], and [15] using RootSIFT) by a large margin [8].

Fig. 1. The Aachen Day-Night dataset contains images taken by hand-held cameras in the old city of Aachen, Germany. These two images were captured during the daytime. The left images demonstrate the same scene in the summer and the right one represents the same square during the winter season. Both images were taken during the daytime.

This dataset was also used to learn keypoint detection and description together with a predictor of the local descriptor discriminativeness. The detection-and-description approach simultaneously outputs sparse, repeatable, and reliable keypoints that outperform state-of-the-art detectors and descriptors on the HPatches dataset and the Aachen Day-Night localization benchmark [9].

[1] https://colmap.github.io/.

2.1 Benchmark for Dataset

For evaluation results, a benchmark is needed. A benchmark can play a role of validity and honesty of results and prevent not so scientifically correct results are considered a new state-of-the-art. On the other hand, methods specializing in the benchmark may be losing generalization power on other benchmarks/datasets.

Because the ECCV conference is one of the significant conferences in the computer vision community, then it is possible to assume the benchmark used during ECCV 2020 Workshop[2] with focus at the Long-Term Visual Localization under Changing Conditions Workshop fulfill the needs for benchmark objectivity and validity.

The submission page[3] for the benchmark allows to submit results to various benchmarking datasets that are commonly used in top-tier computer vision conferences for the problem of long-term visual localization. The benchmarking datasets, that can be evaluated there, include these: Aachen Day-Night and Aachen Day-Night v1.1 datasets, CMU Seasons dataset Localization and Extended CMU Seasons datasets, RobotCar Seasons and RobotCar Seasons v2 datasets, Symphony Seasons dataset, InLoc dataset, and SILDa Weather and Time of Day dataset.

Among many rules and recommendations for benchmark and challenge submission, there are a few which is worth mentioning here:

- Submissions show that existing methods can outperform methods with results published on the benchmarks, e.g., carefully tuning parameters or using a different training dataset.
- Combining
  existing methods with preprocessing or postprocessing approaches, e.g., using histogram equalization on the input images, building better 3D models (for example, through model compression or the use of dense 3D models), or integrating an existing localization algorithm into a (visual) Odometry/SLAM system.

The recommendations of **carefully tuning parameters** and **preprocessing approaches** are used and further expanded in this paper.

3 Used Methods in the Pipeline

In this section, there are briefly mentioned methods that were used in the experiments.

3.1 SuperPoint

The SuperPoint is a self-supervised framework for training interest point detectors and descriptors suitable for many multiple-view geometry problems in computer vision [5]. It is based on a fully convolutional model that operates on

[2] https://sites.google.com/view/ltvl2020/challenges.
[3] https://www.visuallocalization.net.

full-sized images and jointly computes pixel-level interest point locations and associated descriptors in one forward pass. It introduced Homographic Adaptation for boosting interest point detection repeatability. The SuperPoint keypoints and descriptors show excellent performance on the HPatches dataset.

The architecture of SuperPoint is designed as a fully-convolutional neural network architecture that operates on a full-sized image and produces interest point detections accompanied by fixed-length descriptors in a single forward pass.

The architecture is split into two "heads" after the encoder part. Each "head" learns task-specific weights – one for interest point detection and the other one for interest point description.

Most of the network's parameters are shared between these two tasks. That is the difference from traditional systems. The detection and description tasks are handled separately, and the description task needs keypoints locations for its calculation.

The Homographic Adaptation technique was used at training time to improve the base MagicPoint architecture generalization ability on real images. The process can be repeated iteratively to continually self-supervise and improve the interest point detector.

3.2 SuperGlue

SuperGlue [11] is a neural network that matches two sets of local features by jointly finding correspondences and rejecting non-matchable points. Assignments are estimated by solving a differentiable optimal transport problem, whose costs are predicted by a graph neural network.

It introduces a flexible context aggregation mechanism based on attention, enabling SuperGlue to reason about the underlying 3D scene and feature assignments jointly. SuperGlue performs matching in real-time on a modern GPU and can be integrated into modern SfM or SLAM systems.

SuperGlue presents a new way of thinking about the feature matching problem. Instead of trying to learn better features followed by simple heuristic matching (e.g., brute force matching), it learns matching from existing local features using a neural network architecture.

In the context of SLAM, which typically [1] decomposes the problem into the visual feature extraction front-end and the bundle adjustment or pose estimation back-end, the SuperGlue network lies in the middle.

3.3 Hierarchical Multi-scale Attention for Semantic Segmentation

The following semantic segmentation method is described in [18]. Whenever it is referred to the HRNet-OCR in this paper, this modified method is intended, and not the original one from [22].

The main goal of semantic segmentation is to label all pixels within an image as belonging to one of N classes. There are some specific types of predictions

since some are best handled at a lower resolution, and other ones are better handled at a higher resolution.

The multi-scale inference is a commonly used practice for addressing this trade-off. The predictions (at a range of scales) are combined with averaging or max pooling. The results can be improved by using averaging to combine multiple scales. The drawback is the obvious problem of combining the best predictions with poorer ones.

For example, the best prediction for a given pixel comes from the 2× scale, and a much worse prediction comes from the 0.5× scale. Then averaging will combine these predictions, resulting in sub-par output. On the other hand, max-pooling selects only one of N scales to use for a given pixel, while the optimal answer may be a weighted combination across the different scales of predictions.

In multi-scale inference method, there are both the relation and multi-scale context methods [2–4,22] and use multi-scale evaluation to obtain the best results. There are two common approaches used for combining network predictions at multiple scales: average and max pooling. Average pooling involves equally weighting output from different scales which may be sub-optimal. The methods mentioned above share the trait that the network and attention "heads" are trained with a fixed set of scales. Only those scales may be used at run-time, else the network must be retrained.

The recent semantic segmentation work for the Cityscapes dataset has utilized the 20000 coarsely labeled images as-is for training state-of-the-art models [23]. However, a significant amount of each coarse image is unlabelled due to the coarseness of the labels. For achieving state-of-the-art results on the Cityscapes dataset, the method adopts an auto-labeling strategy motivated by [21]. The dense labels are generated for the coarse images in Cityscapes. Most image classification auto-labeling work uses continuous or soft labels but this method generates hard thresholded labels (for storage efficiency and training speed). With soft labels, a teacher network provides a continuous probability for each of N classes for each pixel of an image, whereas for hard labels, a threshold is used to pick a single top-class per pixel. Similar to [6,7] it generates hard and dense labels for the coarse Cityscapes images.

The used auto-labeling method is inspired by recent work on auto-labeling for classification tasks [19,21]. It uses auto-labeling for the Cityscape dataset to improve labeling quality. The Cityscape dataset contains 20000 coarsely labeled images and 3500 finely labeled images. The label quality of the coarse images is very modest and contains a large number of unlabelled pixels. The auto-labeling approach improves label quality.

The soft auto-labeling technique is commonly used. However, it would need to store many labeling data (3.2 TB for the Cityscape dataset). This amount of stored labels would slow the training process considerably.

The method uses the hard labeling strategy. The top class prediction of the teacher network is selected for a given pixel. The labels are based on teacher network output probability and are thresholded. Predictions that exceed the threshold are considered to be correct labels. The threshold is set to 0.9.

4 Evaluation and Ranking

4.1 Evaluation

The pose accuracy is evaluated for each submitted method in each dataset and challenge. The evaluation service follows [12] evaluation approach. It defines a set of thresholds on the position and orientation errors of the estimated pose. For each (X meters, Y degrees) threshold, it reports the percentage of query images localized within X meters and Y degrees of the ground truth pose.

More specifically for the selected Aachen Day-Night dataset, the camera pose estimations thresholds are defined for all conditions (day/night) as follows: the most precise threshold is set to 0.25 m and 2°, the follows threshold is set to 0.50 m and 5°, and the least precise threshold is set to 5.00 m and 10°. These thresholds define how precise the camera pose estimation is, e.g., if a pose estimation fits into the most precise threshold, it is also included in the other two, but when a pose estimation fits only under the least precise threshold, it is not included in the most precise threshold nor middle one.

4.2 Ranking Method

For ranking the methods, it is used the exact same method that was used for the Robust Vision Challenge CVPR 2018[4], Robust Vision Challenge CVPR 2020[5], and in ECCV 2020 Workshop[6].

For the dataset, the submitted results are ranked based on the percentages. The rankings are computed by using the Schulze Proportional Ranking method [17].

The Schulze Proportional Ranking method is based on a pairwise comparison of results. If the results of a method are not available for a dataset, the comparison will assume that it performs worse than a method for which the results are available.

5 Experiments

In this section, there are two main contributions to the state-of-the-art presented. The first contribution outperforms current methods in the selected dataset and the second one confirms that the images can contain information which is not needed for the task of long-term visual localization and can be removed by using semantic segmentation.

Based on selected benchmark in Sect. 2.1 and dataset in Sect. 2 a picking one of the top state-of-the-art performing approaches as a baseline seems to be obvious. One of the top-performing methods on the Aachen benchmarking dataset is the Hierarchical Localization (HLoc) pipeline [10] with Github source

[4] http://www.robustvision.net/rvc2018.php.
[5] http://www.robustvision.net.
[6] https://www.visuallocalization.net/workshop/eccv/2020/.

code[7]. It uses NetVLAD, SuperPoint, SuperGlue, and COLMAP[8]. The used pipeline is described in the paper [10] along with achieved results. The specific parameter settings are mentioned further in this section.

5.1 Parameter Tuning of Hierarchical Localization

If not stated otherwise the parameter was left unchanged and has its value equal to the Original HLoc from the paper [10].

The Aachen Day Images: for experiments with Aachen Day dataset the parameters were set to: NetVLAD: uses 50 most similar images. Image preprocessing: grayscale $True$, resize to 1600 pixels. SuperPoint: 2048 maximum keypoints, keypoint threshold 0.01, NMS radius 3 pixels. SuperGlue: outdoor weights, 100 Sinkhorn iterations. All other parameters were left unchanged and equal to the Original HLoc settings.

The obtained results can be seen in Table 1 in the Day column. The selected parameters helped to outperform previous state-of-the-art methods in the two most precise camera localization thresholds. In the most precise threshold, it was by 0.1% and the next by 0,4%. For the last, the least precise localization threshold, the fine-tuned parameters perform 0.1% worse.

Based on the results, it can be presumed (because ground truth camera poses estimations are unknown, and the evaluation service provides no more information about camera localization) that camera pose estimations for some images were refined (meets sticker thresholds). However, no more new images were localized successfully to pass the least sticker threshold.

The Aachen Night Images: for experiments with the Aachen Night dataset, the HLoc parameters were set to: NetVLAD: uses 50 most similar images. Image preprocessing: grayscale $True$, without any resizing. SuperPoint: 2048 maximum keypoints, keypoint threshold 0.01, NMS radius 3 pixels. SuperGlue: outdoor weights, 100 Sinkhorn iterations. The main difference between the Day and Night setting is that no resize was performed for Night images. All other parameters were left unchanged and equal to the Original HLoc settings.

The obtained results can be seen in Table 1 in the Night column. Results show that the new state-of-the-art results were achieved on the Aachen Night dataset. The improvement was made in the most precise two thresholds (the last one was already at 100% and can not be improved). The previous state-of-the-art results were surpassed by 1.0% in the first threshold and by 1.1% in the second one. Achieved results show that no resizing Aachen Night images can improve localization accuracy.

The Aachen Day-Night Images: the results from Day and Night experiments were combined and named Combined tuned HLoc. The results can be seen in Table 1.

[7] https://github.com/cvg/Hierarchical-Localization.
[8] https://colmap.github.io/.

Table 1. Comparsion of the Original HLoc results in the long-term visual localization benchmark on the Day-Night Aachen dataset compared with Combined tuned HLoc results.

Method	Day [%]	Night [%]
Combined tuned HLoc	**89.7**/**95.8**/98.7	**87.8**/**94.9**/**100.0**
Original HLoc [10]	89.6/95.4/**98.8**	86.7/93.9/**100.0**

5.2 Preprocessing for Semantic Segmentation

The main hypothesis that experiments should prove is that an image from the selected the Aachen dataset contains an area of the image that does not help in the task of long-term visual localization. The area can have associated semantic information from semantic segmentation.

In the case that the area would be removed (masked out and replaced by black color), the selected HLoc pipeline should perform at least as well as Original HLoc or better. If this area is found, it can be stated that the area does not improve the overall results. The moving and dynamic objects in the image could be selected as the area. The area of dynamic objects could be obtained by a semantic segmentation method. The dynamic object could have a high-level interpretation[9] from semantic segmentation classes. Then, groups of classes could be introduced to an aggregate number of classes into a lower number. Furthermore, groups of classes should be used for proving the primary hypothesis.

For long-term visual localization was used HLoc pipeline with the Aachen dataset. For semantic segmentation was used HRNet-OCR method. The performance of HRNet-OCR gives top results[10] in the Cityscape dataset[11] which is similar to the selected Aachen dataset. The implementation of HRNet-OCR was taken from the official Nvidia GitHub[12].

In the task of semantic segmentation multiple different classes can be segmented, e.g. human, bicycle, train, etc. The Cityscape dataset uses the following classes that were grouped into 4 groups. The **Nature** group consist of vegetation and terrain classes. Next, the **sky** group is represented only by sky class. The **Human** group is represented by person and rider classes. And finally, the **vehicle** group aggregates car, truck, bus, caravan, trailer, train, motorcycle, bicycle, and license plate classes. The Cityscape dataset has a few more classes that were omitted because they are rigid or void classes.

[9] For example, car, bus, and truck classes have vehicles (the group of classes) as high-level interpretations.

[10] https://paperswithcode.com/sota/semantic-segmentation-on-cityscapes.

[11] https://www.cityscapes-dataset.com.

[12] https://github.com/NVIDIA/semantic-segmentation.

Several experiments were performed with the Aachen Night dataset, but the current state-of-the-art semantic segmentation methods can not handle the night scenery well. Due to semantic segmentation methods' bad performance during night times, only the Aachen Day images were considered in the next experiments.

At first the semantic segmentations for all Aachen Day images were obtained for example see Fig. 2. From the image that contains all semantic segmentation classes were extracted 4 mentioned groups of classes. The segmented classes in every group were unified and one binary mask was created. The mask was used for masking the original image.

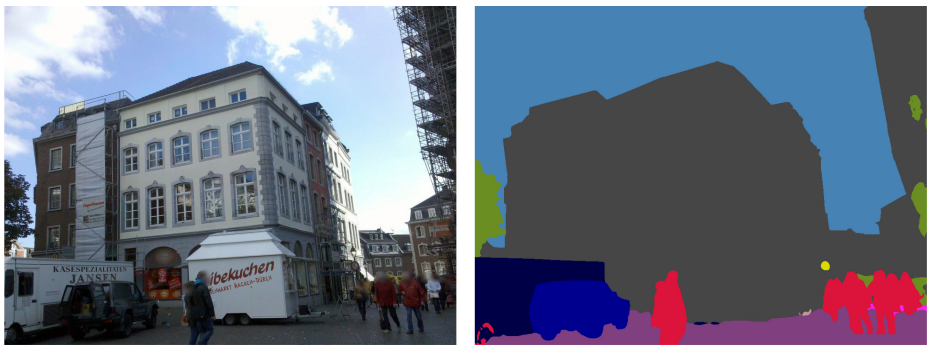

Fig. 2. Left: an original Aachen Day image. Right: semantic segmentation in the color palette of the Cityscape dataset.

All combinations of groups of classes was created, and specific experiments were done. The results can be seen in Table 2. Every experiment was rerun five times and averaged to eliminate randomness. From the table, the top performed result is with the human group of classes and the second best with the vehicle group of classes. Thus, the hypothesis was valid and confirmed.

On the other hand, the experiments that contain the nature group of classes performed poorly and are mostly at the bottom of Table 2. It may be caused because the segmented trees look different in every year season. During Fall, Winter, and Spring seasons, it is possible to see through the deciduous trees (where can buildings be located) and use the visual information that is impossible to see during the Summer season. But semantic segmentation labels not only branches of deciduous trees but also the buildings behind them as a tree class.

Table 2. The combinations of used groups of classes for masking input images of Aachen Day dataset and the precision of localizations.

Semantic Segmentation Groups of Classes	Day [%]
Human	**89.6**/95.5/**98.8**
Vehicle	89.4/95.4/**98.8**
Human, sky, vehicle	89.3/**95.8**/**98.8**
Human, sky	89.3/95.4/**98.8**
Nature, sky	89.2/94.7/98.5
Human, vehicle	89.1/95.4/**98.8**
Sky	89.0/95.4/**98.8**
Sky, vehicle	89.0/95.4/**98.8**
Human, nature, sky	88.6/94.5/98.5
Nature	88.5/94.9/98.5
Human, nature, vehicle	88.5/94.7/98.5
Human, nature, sky, vehicle	88.5/94.7/98.5
Human, nature	88.5/94.4/98.5
Nature, vehicle	88.3/94.8/98.5
Nature, sky, vehicle	88.3/94.5/98.5

6 Conclusion

The final comparison of obtained results is in Table 3 and can be checked on the official benchmark webpage[13]. From the results it can be stated that the experiment with semantic segmentation where a group of classes human was used (named: Masked human HLoc) performs similar to the current state-of-the-art base method Original HLoc. It outperforms Original HLoc by 0.1% in the second-best localization threshold on the Aachen Day dataset and matches results in other localization thresholds. The Combined tuned HLoc method outperforms both Masked human HLoc and Original HLoc methods on the Aachen Day dataset.

Table 3. The comparison of Combined tuned HLoc, Masked human HLoc, and Original HLoc precision.

Method	Day [%]	Night [%]
Combined tuned HLoc	**89.7**/**95.8**/98.7	**87.8**/**94.9**/**100.0**
Masked human HLoc	89.6/95.5/**98.8**	-/-/-
Original HLoc [10]	89.6/95.4/**98.8**	86.7/93.9/**100.0**

[13] https://www.visuallocalization.net/benchmark/.

In the future work we will perform experiments with semantic segmentation on other datasets and we will examine more semantic segmentation approaches that can lead us to more precise segmentation and thus enhance long-term visual localization precision.

Acknowledgements. This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

1. Cadena, C., et al.: Past, present, and future of simultaneous localization and mapping: towards the robust-perception age. IEEE Trans. Rob. **32**(6), 1309–1332 (2016)
2. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR abs/1706.05587 (2017)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, February 2018
4. Cheng, B., et al.: Panoptic-deeplab: a simple, strong, and fast baseline for bottom-up panoptic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 12472–12482. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01249
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: self-supervised interest point detection and description. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2018
6. Lee, D.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks (2013)
7. Liu, L., Li, Y., Tan, R.T.: Decoupled certainty-driven consistency loss for semi-supervised learning (2020)
8. Widya, A.R., Torii, A., Okutomi, M.: Structure from motion using dense CNN features with keypoint relocalization. IPSJ Trans. Comput. Vis. Appl. **10**(1), 1–7 (2018). https://doi.org/10.1186/s41074-018-0042-y
9. Revaud, J., et al.: R2D2: repeatable and reliable detector and descriptor. CoRR abs/1906.06195 (2019)
10. Sarlin, P., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: robust hierarchical localization at large scale. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12708–12717 (2019). https://doi.org/10.1109/CVPR.2019.01300
11. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: learning feature matching with graph neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
12. Sattler, T., et al.: Benchmarking 6DOF outdoor visual localization in changing conditions. In: Conference on Computer Vision and Pattern Recognition, pp. 8601–8610 (2018)
13. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: British Machine Vision Conference, September 2012
14. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

15. Schönberger, J.L., Price, T., Sattler, T., Frahm, J.M., Pollefeys, M.: A vote-and-verify strategy for fast spatial verification in image retrieval. In: Asian Conference on Computer Vision (ACCV) (2016)
16. Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 501–518. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_31
17. Schulze, M.: A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. Soc. Choice Welfare **36**, 267–303 (2011). https://doi.org/10.1007/s00355-010-0475-4
18. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation (2020)
19. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 1195–1204 (2017)
20. Toft, C., et al.: Long-term visual localization revisited. IEEE Trans. Pattern Anal. Mach. Intell. 14 (2020)
21. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
22. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12351, pp. 173–190. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_11
23. Yuan, Y., Wang, J.: OCNet: object context network for scene parsing. CoRR abs/1809.00916 (2018)